FacTA: Evaluation of Event Factuality and Temporal Anchoring

Anne-Lyse Minard, Manuela Speranza, Rachele Sprugnoli and Tommaso Caselli

DOI: 10.4000/books.aaccademia.1509
Publisher: Accademia University Press
Place of publication: Torino
Year of publication: 2015
Published on OpenEdition Books: 11 November 2016
Series: Collana dell'Associazione Italiana di Linguistica Computazionale
Electronic ISBN: 9788899200008

Electronic reference
FacTA: Evaluation of Event Factuality and Temporal Anchoring

Anne-Lyse Minard\textsuperscript{1}, Manuela Speranza\textsuperscript{1}, Rachele Sprugnoli\textsuperscript{1,2}, Tommaso Caselli\textsuperscript{3}

\textsuperscript{1}Fondazione Bruno Kessler, Trento
\textsuperscript{2}Università di Trento
\textsuperscript{3}VU Amsterdam
{minard, manspera, sprugnoli}@fbk.eu
\vspace{1pt}
t.caselli@vu.nl

Abstract

\textbf{English.} In this paper we describe FacTA, a new task connecting the evaluation of factuality profiling and temporal anchoring, two strictly related aspects in event processing. The proposed task aims at providing a complete evaluation framework for factuality profiling, at taking the first steps in the direction of narrative container evaluation for Italian, and at making available benchmark data for high-level semantic tasks.

\textbf{Italiano.} Questo articolo descrive FacTA, un nuovo esercizio di valutazione su fattualità ed ancoraggio temporale, due aspetti dell’analisi degli eventi strettamente connessi tra loro. Il compito proposto mira a fornire una cornice completa di valutazione per la fattualità, a muovere i primi passi nella direzione della valutazione dei contenitori narrativi per l’italiano e a rendere disponibili dati di riferimento per compiti semantici di alto livello.

1 Introduction

Reasoning about events plays a fundamental role in text understanding; it involves different aspects, such as event identification and classification, temporal anchoring of events, temporal ordering, and event factuality profiling. In view of the next EVALITA edition (Attardi et al., 2015),\textsuperscript{1} we propose FacTA (\textit{F}actuality a\textit{n}d \textit{e}\textit{r}\textit{t}\textit{m}\textit{a}\textit{p}\textit{i}\textit{a}l \textit{a}n\textit{c}h\textit{o}r\textit{i}ng), the first task comprising the evaluation of both factuality profiling and temporal anchoring, two strictly interrelated aspects of event interpretation.

Event factuality is defined in the literature as the level of committed belief expressed by relevant sources towards the factual status of events mentioned in texts (Saurí and Pustejovsky, 2012). The notion of factuality is closely connected to other notions thoroughly explored by previous research conducted in the NLP field, such as subjectivity, belief, hedging and modality; see, among others, (Wiebe et al., 2004; Prabhakaran et al., 2010; Medlock and Briscoe, 2007; Saurí et al., 2006). More specifically, the factuality status of events is related to their degree of certainty (from absolutely certain to uncertain) and to their polarity (affirmed vs. negated). These two aspects are taken into consideration in the factuality annotation frameworks proposed by Saurí and Pustejovsky (2012) and van Son et al. (2014), which inspired the definition of factuality profiling in FacTA.

Temporal anchoring consists of associating all temporally grounded events to time anchors, i.e. temporal expressions, through a set of temporal links. The TimeML annotation framework (Pustejovsky et al., 2005) addresses this issue through the specifications for temporal relation (TLINK) annotation, which also implies the ordering of events and temporal expressions with respect to one another. Far from being a trivial task (see systems performance in English (UzZaman et al., 2013) and in Italian (Mirza and Minard, 2014)), TLINK annotation requires the comprehension of complex temporal structures; moreover, the number of possible TLINKs grows together with the number of annotated events and temporal expressions. Pustejovsky and Stubbs (2011) introduced the notion of \textit{narrative container} with the aim of reducing the number of TLINKs to be identified in a text while improving informativeness and accuracy.

A narrative container is a temporal expression or an event explicitly mentioned in the text into which other events temporally fall (Styler IV et al., 2014). The use of narrative containers proved to be useful to accurately place events on timelines in the domain of clinical narratives (Miller et al., 2013). Temporal anchoring in FacTA moves in the direction of this notion of narrative container by fo-

\textsuperscript{1}http://www.evalita.it/
cusing on specific types of temporal relations that link an event to the temporal expression to which it is anchored. However, anchoring events in time is strictly dependent of their factuality profiling. For instance, counterfactual events will never have a temporal anchor or be part of a temporal relation (i.e. they never occurred); this may not hold for speculated events, whose association with a temporal anchor or participation in a temporal relation is important to monitor future event outcomes.

2 Related Evaluation Tasks

Factuality profiling and temporal anchoring of events are crucial for many NLP applications (Wiebe et al., 2005; Karttunen and Zaenen, 2005; Caselli et al., 2015) and therefore have been the focus, either direct or indirect, of several evaluation exercises, especially for English.

The ACE Event Detection and Recognition tasks of 2005 and 2007 (LDC, 2005) took into consideration factuality-related information by requiring systems to assign the value of the modality attribute to extracted events so as to distinguish between asserted and non-asserted (e.g. hypothetical, desired, and promised) events. Following the ACE evaluation, a new task has recently been defined in the context of the TAC KBP 2015 Event Track. The Event Nugget Detection task aims at assessing the performance of systems in identifying events and their realis value, which can be ACTUAL, GENERIC or OTHER (Mitamura et al., 2015). Other tasks focused on the evaluation of speculated and negated events in different domains such as biomedical data and literary texts (Nédellec et al., 2013; Morante and Blanco, 2012).

The evaluation of event modality was part of the Clinical TempEval task at SemEval 2015 (Bethard et al., 2015), which also proposed for the first time the evaluation of narrative container relations between events and/or temporal expressions.

Temporal anchoring has been evaluated in the more general context of temporal relation annotation in the 2007, 2011 and 2013 TempEval evaluation exercises (Verhagen et al., 2007; Verhagen et al., 2010; Uzzaman et al., 2013) as well as in the EVENTI task (Caselli et al., 2014) on Italian at EVALIT.A 2014. The TimeLine task at SemEval 2015 (Minard et al., 2015) was the first evaluation exercise focusing on cross-document event ordering; in view of the creation of timelines, it requires temporal anchoring and ordering of certain and non-negated events.

With respect to the aforementioned tasks, FacTA aims at providing a complete evaluation framework for factuality profiling, at taking the first steps in the direction of narrative container evaluation for Italian, and at making new datasets available to the research community.

3 Task Description

The FacTA task consists of two subtasks: factuality profiling and temporal anchoring of given gold event mentions. Participants may decide to take part to both or only one of the proposed subtasks.

3.1 Subtask 1: Factuality Profiling

Tonelli et al. (2014) propose an annotation schema of factuality for English based on the annotation framework by van Son et al. (2014). This schema was then adapted to Italian by Minard et al. (2014). Following this, we represent factuality by means of a combination of three attributes associated with event mentions: certainty, time, and polarity. For each given gold event mention, participant systems are required to assign values for three factuality attributes.

The certainty attribute relates to how sure the main source is about the mentioned event and admits the following four values: certain, possible, probable, and underspecified.

The time attribute specifies the time when an event is reported to have taken place or to be going to take place. Its values are non_future (for present and past events), future (for events that will take place), and underspecified.

The polarity attribute captures if an event is affirmed or negated and, consequently, it can be either positive or negative; when there is not enough information available to detect the polarity of an event mention, it is underspecified. The event’s annotation framework, inspired by FactBank (Saurí and Pustejovsky, 2009), enriches it with the distinction between future and non-future events.

The main source is either the utterer (in direct speech, indirect speech or reported speech) or the author of the news (in all other cases). In this framework, where factuality depends strictly on the source, factuality annotation is also referred to as attribution annotation.

4van Son et al.’s annotation framework, inspired by FactBank (Saurí and Pustejovsky, 2009), enriches it with the distinction between future and non-future events.

5The main source is either the utterer (in direct speech, indirect speech or reported speech) or the author of the news (in all other cases). In this framework, where factuality depends strictly on the source, factuality annotation is also referred to as attribution annotation.
Factuality value. The combination of the attributes described above determines the value of an event: factual, counterfactual or non_factual. More specifically, the overall factuality value is factual if its values are certain, non_future, and positive (e.g. ‘rassegna’ in [1]), while it is counterfactual (i.e. the event is reported as not having taken place) if its values are certain, non_future, and negative (e.g. ‘nominato’ in [2]). In any other combination, the event is non factual, either because it is non certain, or future (e.g. ‘nomineranno’ in [1]).

(1) Smith ha rassegnato ieri le dimissioni; nomineranno il suo successore entro un mese. (“Smith resigned yesterday; they will appoint his replacement within a month.”)

(2) Non ha nominato un amministratore delegato. (“He did not appoint a CEO.”)

No factuality annotation. Language is used to describe events that do not correlate with a real situation in the world (e.g. ‘parlare’ in [3]). For these event mentions participant systems are required to leave the value of all three attributes empty.

(3) Guardate, penso che sia prematuro parlare del nuovo preside (“Well, I think it is too early to talk about the new dean”)

3.2 Subtask 2: Temporal Anchoring

Given a set of gold events, participant systems are required to detect those events for which it is possible to identify a time anchor. Our definition of time anchor includes two different types of elements: the temporal expressions occurring in the text, as well as the Document Creation Time (DCT), which is part of the metadata associated with each document. The subtask thus includes temporal expression (or TIMEX3) detection and normalization, as well as identification of temporal relations (or TLINKs) between events and temporal expressions.

TIMEX3 detection and normalization. Based on the annotation guidelines produced within the NewsReader project (Tonelli et al., 2014), which in turn are based on the ISO-TimeML guidelines (ISO TimeML Working Group, 2008), this consists of:

- TIMEX3 detection: identification and classification of temporal expressions of type date and time (durations and sets of times, on the other hand, are excluded from the task).
- TIMEX3 normalization: identification of the value attribute for each temporal expression.

For instance, in [1], ieri is a TIMEX3 of type date with value 2015-07-28 considering 2015-07-29 as DCT.

TLINK identification. This consists of detecting TLINKs of types IS_INCLUDED and SIMULTANEOUS holding between an event and a TIMEX3 (i.e. the anchor of the event), as defined in (Tonelli et al., 2014). The event (the source of the TLINK) and the TIMEX3 (the target) can either appear in the same sentence or in different sentences. For instance, in [1], rassegnato is anchored to ieri (rassegna, IS_INCLUDED, ieri).

4 Dataset Description

4.1 Subtask 1: Factuality Profiling

As a training dataset, participants can use Fact-Ita Bank (Minard et al., 2014), which consists of 170 documents selected from the Ita-TimeBank (Caselli et al., 2011), which was first released for the EVENTI task at EVALITA 2014. Fact-Ita Bank contains annotations for 10,205 event mentions and is already distributed with a CC-BY-NC license.

System evaluation will be performed on the “first five sentences” section of WItaC, the NewsReader Wikinews Italian Corpus (Speranza and Minard, 2015). It consists of 15,676 tokens and has already been annotated with event factuality (as this annotation has been projected from English, it will need some minor revision).

4.2 Subtask 2: Temporal Anchoring

For temporal expression detection and normalization, participant systems can be trained on the dataset used for the EVENTI Task at Evalita 2014 (Caselli et al., 2014). It also contains TLINKs between events and TIMEX3s in the same sentence but not in different sentences. To make it usable as a training corpus for temporal anchoring, we would have to add the TLINKs between events and

---

---
TIMEX3s in different sentences and the TLINKs between events and the DCT, which would require a big effort. Thus, we are instead planning to add the needed relations to only a subset of the corpus, namely the same 170 documents that compose Fact-ItaBank.

As test data we will use the “first five sentences” section of WItaC (Speranza and Minard, 2015), which is already annotated with TIMEX3s and with TLINKs between events and TIMEX3s in the same sentences; the test set thus needs to be completed through the addition of TLINKs between events and TIMEX3s in different sentences.

5 Evaluation

Each subtask will be evaluated independently. No global score will be computed as the task aims to isolate the two phenomena.

5.1 Subtask 1: Factuality Profiling

Participant systems will be evaluated in terms of precision, recall and their harmonic mean (i.e. F1 score). We will perform the evaluation of:

- values of the factuality attributes (polarity, certainty and time);
- detection of events to which factuality values should not be assigned (i.e. “no factuality annotation” events);
- assignment of the overall factuality value (combination of the three attributes), including also the non-assignment of factuality attributes.

The official ranking of the systems will be based on the evaluation of the overall factuality value.

5.2 Subtask 2: Temporal Anchoring

For the temporal anchoring subtask, we will evaluate the number of event-TIMEX3 relations correctly identified in terms of precision, recall and F1 score. Two relations in the reference and the system prediction match if their sources and their targets match. Two sources (i.e. events) are considered as equivalent if they have the same extent, whereas two targets (i.e. TIMEX3s) match if their values are the same. Participant systems will be ranked according to the F1 score.

We will not apply the metric for evaluating temporal awareness based on temporal closure graphs proposed by UzZaman and Allen (2011), which is unnecessarily complex as we have reduced the relations to only IS_INCLUDED and SIMULTANEOUS.

6 Discussion and Conclusions

The FacTA task connects two related aspects of events: factuality and temporal anchoring. The availability of this information for Italian will both promote research in these areas and fill a gap with respect to other languages, such as English, for a variety of semantic tasks.

Factuality profiling is a challenging task aimed at identifying the speaker/writers degree of commitment to the events being referred to in a text. Having access to this type of information plays a crucial role for distinguishing relevant and non-relevant information for more complex tasks such as textual entailment, question answering, and temporal processing.

On the other hand, anchoring events in time requires to interpret temporal information which is not often explicitly provided in texts. The identification of the correct temporal anchor facilitates the organization of events in groups of narrative containers which could be further used to improve the identification and classification of in-document and cross-document temporal relations.

The new annotation layers will be added on top of an existing dataset, the EVENTI corpus, thus allowing to re-use existing resources and to promote the development of multi-layered annotated corpora; moreover a new linguistic resource, WItaC, will be provided. The availability of these data is to be considered strategic as it will help the study the interactions of different language phenomena and enhance the development of more robust systems for automatic access to the content of texts. The use of well structured annotation guidelines grounded both on official and de facto standards is a stimulus for the development of multilingual approaches and promote discussions and reflections in the NLP community at large.

Considering the success of evaluation campaigns such as Clinical TempEval at SemEval 2015 and given the presence of an active community focused on extra-propositional aspects of meanings (e.g. attribution11), making available new annotated data in the framework of an evaluation campaign for a language other than English can have a large impact in the NLP community.

11 Ex-Prom Workshop at NAACL 2015 http://www.cse.unt.edu/exprom2015/
Acknowledgements

This work has been partially supported by the EU NewsReader Project (FP7-ICT-2011-8 grant 316404) and the NWO Spinoza Prize project Understanding Language by Machines (sub-track 3).

References


